International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4
ISSN: 2454-4248
202 – 206

Document Indexing Strategies in Big Data

A Survey

K.Swapnika
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
swapnika.griet@gmail.com

K.Swanthana
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
swanthana.griet@gmail.com

Y.Krishna Bhargavi
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
kittu.bhargavi@gmail.com

***Abstract*—** From past few years, the operations of the Internet have a significant growth and individuals, organizations were unaware of this data explosion. Because of the increasing quantity and diversity of digital documents available to end users, mechanism for their effective and efficient retrieval is given highest importance. One crucial aspect of this mechanism is indexing, which serves to allow documents to be located quickly. The problem is that users want to retrieve on the basis of context, and individual words provide unreliable evidence about the contextual topic or meaning of a document. Hence, the available solutions cannot meet the needs of the growing heterogeneous data in terms of processing. This results in inefficient information retrieval or search query results. The design of indexing strategies that can support this need is required. There are various indexing strategies which are utilized for solving Big Data management issues, and can also serve as a base for the design of more efficient indexing strategies. The aim is to explore document indexing strategy for Big Data manageability. The existing systems like, Latent Semantic Indexing , Inverted Indexing, Semantic indexing and Vector Space Model has their own challenges such as, Demands high computational performance, Consumes more memory Space, Longer data processing time, Limits the search space, will not produce the exact answer, Can present wrong answers due to synonyms and polysemy, approach makes use of formal ontology. This paper will describe and compare the various Indexing techniques and presents the characteristics and challenges involved.

Keywords*-Big Data, Indexing, Query, Retrieval, Contextual, Heterogeneous*

*****

I. INTRODUCTION

Collection of the digital information in terms of structured or unstructured data is known as big data. There has been an extraordinary growth in the amount of digital information available to the common user. The factors that contributed to this growth include the world-wide proliferation of the Internet, the economical cost of digitizing information contents into computerized forms, and a general increase in computer literacy and accessibility to a large and growing number of people around the world. Because of the increasing quantity and diversity of digital documents available to end users, mechanism for their effective and efficient retrieval is given importance. One crucial aspect of this mechanism is indexing, which serves to allow documents to be located quickly. Building an index for textual documents traditionally contains selected terms, where each term indicates the locations where it occurs. Without an index, the system must examine each and every available document in the repository to determine whether it contains the term. With an index, however, the system simply searches through the index data structure for particular word to identify and locate the documents containing it.

Traditionally, indexing is a manual task done by professional indexers where they have the flexibility in choosing which terms to use in the indexes. However, the large volume of digitized text documents has necessitated the use of automated indexing by computer systems. In automatic indexing, algorithms are used to determine which terms to use in the indexes. For instance, a content-bearing term that occurs frequently within a document is more likely to be important in that document and should be included as an index term.

Indexing techniques have been developed in order to make possible the identification of the information content in documents. In general, indexes permit the representation of knowledge about a domain in order to facilitate access to information. It simply means pointing to or indicating the content, meaning, purpose and features of messages, texts and documents. Indexes are a list of tags, names, subjects, etc. of a group of items which references where data can be found. Typically the indexing of a textual document is obtained through the identification of a set of terms or keywords which characterize the document content (i.e terms) which describe the topics dealt with in the document. The terms included in this set have not only to be representative of the topics covered in the documents, but they also need to be distinguishing, in

 ISSN: 2454-4248
202 – 206

that they should make it possible to discriminate one document against the other documents in the collection covering the same or similar topics. An indexing strategy is the design of an access method to a searched item. It also describes how data is organized in a storage system.

Big Data indexing depends on a solution that utilizes a massively parallel computer or machine that interconnects lots of RAM, CPUs, and disk units [1]. The benefits of this are high throughput for data processing, decreased access time for queries, data replication that results in increased availability and reliability, and scalability of the structure. The design of an access method or the type of indexing strategy to be used in processing a specific dataset depends on the type of queries that processed on the dataset, such as similarity queries (nearest neighbor search), range queries, point query, keyword queries, and ad-hoc query. Therefore, it is must to be aware of the type of data to be indexed (e.g. logs, email, audio, video, images, etc.) and the type of query that will be performed on the indexes.

II. RELATED WORK

Today the web is the main source for the text documents, the amount of textual data available to us is consistently increasing, and approximately 80% of the information of an organization is stored in unstructured textual format [2], in the form of reports, email, views and news etc. The [3] shows that approximately 90% of the world's data is held in unstructured formats, so Information intensive business processes demand that we transcend from simple document retrieval to knowledge discovery. The need of automatically retrieval of useful knowledge from the huge amount of textual data in order to assist the human analysis is fully apparent [4]. Market trend based on the content of the online news articles, sentiments, and events is an emerging topic for research in data mining and text mining community [5]. For these purpose state-of-the-art approaches to text classifications are presented in [6], in which three problems were discussed: documents representation, classifier construction and classifier evaluation. So constructing a data structure that can represent the documents, and constructing a classifier that can be used to predicate the class label of a document with high accuracy, are the key points in text classification.

One of the purposes of research is to review the available and known work, so an attempt is made to collect what's known about the documents classification and representation. This paper covers the overview of syntactic and semantic matters, domain ontology and tokenization concern and focused on the different machine learning techniques for text classification using the existing literature. Information Extraction (IE) methods is aim to extract specific information from text documents. This is the first approach assumes that text mining essentially corresponds to information extraction.

In this paper we have used system literature review process and followed standard steps for analyzing different document indexing approaches. First of all we tried to search for relevant papers, presentations, research reports and policy documents that were broadly concerned with documents classification and Indexing. Potentially relevant papers were identified using the electronic databases and websites, Such as IEEE Explore, Springer Linker, Science Direct, ACM Portal and Googol Search Engine. For best and consistent search a systematic search strategy was adopted. Through the above process we have identified some of the Indexing strategies, hence categorized and these approaches are studied in detail for comparison of their methodology. In below section we will discuss those approaches and their challenges that are to be overcome.

III. METHODS

Indexing in big data is to speed up the data retrieval process and to minimize the search query processing time. According to [2], indexing strategies can be categorized into Artificial Intelligence (AI) approach, and Non-Artificial Intelligence (NAI) approach.

A. Artificial Intelligence (AI) Approachs

It is which has the ability to detect the unknown behavior of data by observing the patterns and categorizing the data using its knowledge base. It consumes more time for retrieval as data gets continuously changing and frequently gets updated, it maintains good accuracy.

It is further sub-divided into two techniques 1. Latent Semantics. 2. Hidden Markov Model

1) Latent Semantic Indexing [LSI]: Latent Semantic Indexing, LSI for short, is an indexing strategy (retrieval/access method) that identifies patterns between the terms in an unstructured data set (specifically, text)[6]. It uses a mathematical approach known as Singular Value Decomposition (SVD) [7] for the pattern or relationship identification.

The first step is to represent the text as a matrix in which each row stands for a unique word and each column stands for a text passage or other context. Each cell contains the frequency with which the word of its row appears in the passage denoted by its column. LSA applies singular value decomposition (SVD) to the term-document matrix. This is a form of factor analysis. In SVD, a rectangular matrix is decomposed into the product of three other matrices: One component matrix describes the original row entities as vectors of derived orthogonal factor values, second describes the original column entities in the same way, and the third is a diagonal matrix containing scaling values such that when the three components are matrix-multiplied, the original matrix is reconstructed [5].

 ISSN: 2454-4248
202 – 206

The challenges mostly faced while working with LSI is that the model obtained from truncated SVD which is costlier, scalability and performance. LSI strategy demands very high computational performance as well as memory to index Big Data. LSI [9,10] supports keyword queries on textual data which can be in the form of web contents (images, audio, etc.), documents, emails, or any item that can be converted into text.

2) Hidden Markov Model [HMM]: Hidden Markov Model is an indexing approach which is developed from the Markov model [5]. It is consists of states which are associated by transitions, where future states are completely dependent on present state and independent on historical states. In this technique query results are generally predictions of future states of an item, based on the current state. The present state is used to predict the future states using the dependent data which increases a good performance.

The HMM is a sequence model [6]. A sequence model or sequence classifier is a model whose job is to assign a label or class to each unit in a sequence, thus mapping a sequence of observations to a sequence of labels. An HMM is a probabilistic sequence model: given a sequence of units (words, letters, morphemes, sentences, whatever), they compute a probability distribution over possible sequences of labels and choose the best label sequence. Sequence labeling tasks come up throughout speech and language processing, a fact that isn't too surprising if we consider that language consists of sequences at many representational levels. These include part-of-speech tagging named entity tagging, and speech recognition among others.

B. Non – Artificial Intelligence Approachs

In NAI indexing approach, the formation of indexes does not depend on the meaning of the data item or the relationship between texts. Rather, indexes are formed based on items most queried or searched for in a particular data set. The inverted indexing approach is an NAI indexing approach.

1) Inverted Indexing

In its basic form, an inverted index consists of postings lists, one associated with each term that appears in the collection. The structure of an inverted index is illustrated in Figure below. A postings list is comprised of individual postings, each of which consists of a document id and a payload information about occurrences of the term in the document. The simplest payload is nothing, For simple boolean retrieval, no additional information is needed in the posting, other than the document id; the existence of the posting itself indicates that presence of the term in the document. The most common payload, however, is term frequency [5], or the number of times the term occurs in the document. More complex payloads include positions of every occurrence of the term in the document, properties of the term, or even the results of additional linguistic processing [8].In the web context, anchor text information is useful in enriching the representation of document content; this information is often stored in the index as well.

2) Semantic indexing

The goal of semantic indexing [9] is to use semantic information to improve the quality of information retrieval. While using purely lexical approach to indexing is adequate in most situations, there are significant limitations. For example, suppose a user wants to do a search on portable computers, he would go on to the system and issue a query to search for documents containing the term "notebook". The problem is that if the index is strictly keyword-based, then documents containing the term "laptop" would not be included in the search results. This is a fundamental problem of synonymy where two different words describe the same concept. To make matter worse, documents related to writing stationery (paper notebook) would also be included in the search results even though they are not relevant to the user's need. This is a fundamental problem of polysemy where the same word has multiple conceptual meanings.

When the user issues the query "notebook", what he actually wants is not documents containing the word "notebook"; rather, he wants documents pertaining to or about the concept of notebook computers. The inability for the system to understand user's intention leads to low-quality search results. As a consequence, the user must rely on manual search strategies obtained through experience to improve search quality. For example, he may issue multiple queries using synonyms or he may use exclusion terms to filter irrelevant search results.

Hence, in order to improve search quality and make searching an easier task for users, there needs to be an integration of semantics (human meaning) into the process of document indexing. One simple approach is to use a thesaurus, which contains a list of terms and their corresponding synonyms, during the indexing process. However, this approach does not really solve the problem because, based on keywords alone, the system still does not know whether the user actually means a computer notebook or a writing stationery notebook. It is quite evident that contextual information is crucial in determining what the user means.

The problem of determining the intended meaning of users' queries is a difficult one and has many facets to it. How does the system determine the context of a query? Similarly, how does the user specify the query context to the system? The question essentially becomes: how do the system and the user agree. This approach makes use of formal ontology.

3) Vector Space Model

 ISSN: 2454-4248
202 – 206

The vector-space model [9,10] is based on the representation of both documents and queries as weighted vectors in the space of the index terms, whose dimensionality is determined by the size of the vocabulary used in the indexing process. A similarity measure is used to cluster together documents which show the higher degree of similarity. A vector-based information retrieval method represents both documents and queries with high-dimensional vectors, while computing their similarities by the vector inner product. When the vectors are normalized to unit lengths, the inner product measures the cosine of the angle between the two vectors in the vector space. Once the terms have been associated with weights, documents can be represented by term vectors.

IV. COMPARISION OF INDEXING STRATEGIES

TABLE I will summarize the various types of indexing strategies, along with the possible type of data and queries they support. Table II outlines the main characteristics of each indexing strategy. The key features and challenges faced in each are described.

TABLE I. INDEXING STRATEGIES AND QUERY-TYPES

Indexing Strategies	***Type of Input***	***Query-type***
Inverted Indexing	Multimedia data, Documents	Keyword search
LSI	Multimedia data, spatial data (textual data)	Keyword search
HMM	Speech, Documents, passages	Keyword search
Semantic Indexing	Documents, textual data	Concept search
Vector space model	Documents	Keyword search

TABLE II. CHARECTERISTICS AND CHALLENGES OF INDEXING STRATEGIES

Indexing Strategies	***Strategies***	***Properties Challenges***
Inverted Indexing	- Index consumes less space - Full text search (keyword search)	- Longer data processing time - Limits the search space, will not produce the exact answer - Can present wrong answers due to synonyms and polysemy.
LSI	LSI - Uses data and meaning of data for indexing - Presents accurate query results (since it uses more information)	- Demands high computational performance - Consumes more memory Space
HMM	-Uses position of occurrences as index - allows the efficient modeling of the order of the keywords in a query.	- Probabilities of associating them to different database terms. - training collections for automatic optimization of weights for arbitrary features
Semantic Indexing	SI- It is highly adaptable to patterns of usage.	- This approach makes use of formal ontology. -Here, the domain is described by controlled vocabulary.
Vector space model	VSM- it represents both documents and queries with high-dimensional vector - measures the cosine of the angle between the two vectors in the vector space for classification.	-There is no real theoretical basis for the assumption of a term space. -it is more visualized. - Terms are not independent of all other terms.

 ISSN: 2454-4248
202 – 206

V. CONCLUSION

In this paper, we conclude that there is a requirement of some approach such that it solves some of the challenges that are noticed in the existing techniques. Hence, in future there is a need to implement a novel document indexing technique using deep learning algorithms which supports and optimizes the memory consumption, lower the data processing time and resulting in enhanced Performance of the indexing using big data analytics.

REFERENCES

[1] Adamu, Fatima Binta, et al. A Survey On Big Data Indexing Strategies. No. SLAC-PUB-16460. SLAC National Accelerator Laboratory (SLAC), 2016.

[2] A. Gani, A. Siddiqa, S. Shamshirband, and F. Hanum, "A survey on indexing techniques for big data: taxonomy and performance evaluation,"Knowledge and Information Systems, pp. 1–44, 2015.

[3] Vaishanvi Gawande, Ambiks Pawar, "Survey on Big Data Indexing Techniques" International Science Press, pp. 619-624, ISSN: 0974-5572,IJCTA, Aug 2017.

[4] Trong Nhan Phan , Markus Jäger , Stefan Nadschläger , Josef Küng , and Tran Khanh Dang "An Efficient Document Indexing-Based Similarity Search in Large Datasets" © Springer International Publishing Switzerland 2015 T.K. Dang et al. (Eds.): FDSE 2015, LNCS 9446, pp. 16–31, 2015. DOI: 10.1007/978-3-319-26135-5_2.

[5] Ekta Chauhan , Dr. Amit Asthana, "Review of Indexing Techniques in Information Retrieval" International Journal of Engineering Science and Computing, Research Article, Volume 7 Issue No.7, July 2017.

[6] Widodo and W. Wibowo, "Improving classification performance by extending documents terms," in Data and Software Engineering (ICODSE), 2014 International Conference on, pp. 1–5, Nov 2014.

[7] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis", Journal of the American Society for Information Science (1986-1998); Sep 1990.

[8] Ajit Kumar Mahapatra , Sitanath Biswas, "Inverted indexes: Types and techniques " IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011 ISSN (Online): 1694-0814.

[9] Marco Suárez Barón* , Kathleen Salinas Valencia An approach to semantic indexing and information retrieval, Rev. Fac. Ing. Univ. Antioquia N.° 48. pp. 165-187. June, 2009

[10] Ch. Aswani Kumar, Ankush Gupta, Mahmooda Batool and Shagun Trehan, "Latent Semantic Indexing-Based Intelligent Information Retrieval System for Digital Libraries" Journal of Computing and Information Technology - CIT 14, 2006, 3, 191–196 doi:10.2498/cit.2006.03.02.

[11] Yu, Kai, Shipeng Yu, and Volker Tresp. "Multi-label informed latent semantic indexing." Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005.

[12] S. Deerwester, S. T. Dumais, T. K. Landauer. "Indexing by latent semantic analysis". Journal of the American Society of Information Science. Vol. 41. 1990. pp.391- 407.